"You Know What to Do": Proactive Detection of YouTube Videos Targeted by Coordinated Hate Attacks*

Enrico Mariconti
UCL
e.mariconti@ucl.ac.uk

Guillermo Suarez-Tangil
King's College London
guillermo.suarez-tangil@kcl.ac.uk

Jeremy Blackburn
Binghamton University
blackburn@cs.binghamton.edu

Emiliano De Cristofaro
UCL & Alan Turing Institute
e.decristofaro@ucl.ac.uk

Nicolas Kourtellis
Telefonica Research
nicolas.kourtellis@telefonica.com

Ilias Leontiadis
Samsung AI
i.leontiadis@samsung.com

Jordi Luque Serrano
Telefonica Research
jordi.luqueserrano@telefonica.com

Gianluca Stringhini
Boston University
gian@bu.edu

Abstract

Video sharing platforms like YouTube are increasingly targeted by aggression and hate attacks. Prior work has shown how these attacks often take place as a result of *"raids,"* i.e., organized efforts by ad-hoc mobs coordinating from third-party communities. Despite the increasing relevance of this phenomenon, however, online services often lack effective countermeasures to mitigate it. Unlike well-studied problems like spam and phishing, coordinated aggressive behavior both targets and is perpetrated by humans, making defense mechanisms that look for automated activity unsuitable. Therefore, the de-facto solution is to reactively rely on user reports and human moderation.

In this paper, we propose an automated solution to identify YouTube videos that are likely to be targeted by coordinated harassers from fringe communities like 4chan. First, we characterize and model YouTube videos along several axes (metadata, audio transcripts, thumbnails) based on a ground truth dataset of videos that were targeted by raids. Then, we use an ensemble of classifiers to determine the likelihood that a video will be raided with very good results (AUC up to 94%). Overall, our work provides an important first step towards deploying proactive systems to detect and mitigate coordinated hate attacks on platforms like YouTube.

1 Introduction

Over the years, the Web has shrunk the world, allowing individuals to share viewpoints with many more people than they are able to in real life. At the same time, however, it has also enabled anti-social and toxic behavior to occur at an unprecedented scale. As social interactions increasingly take place online, cyber-aggression has unfortunately become a pressing problem [34]. In particular, coordinated harassment campaigns are more and more frequent, with perpetrators working together to repeatedly target victims with hateful comments [11, 15, 24]. One example of such behavior is a phenomenon known as *raiding*, whereby ad-hoc mobs coordinate on social platforms to organize and orchestrate attacks aimed to disrupt other platforms and undermine users who advocate for issues and policies they do not agree with [35, 44].

Abusive activity is generated by humans and not by automated programs, thus, systems to detect unwanted content/bots [7, 53, 71] are not easily adapted to this problem. In fact, Google's CEO, Sundar Pichai, recently identified detecting hate speech as one of the most difficult challenges he is facing [78]. Hence, platforms mostly adopt *reactive* solutions, letting users report abusive accounts and taking actions according to terms of services, e.g., blocking or suspending offenders [43]. However, this approach is inherently slow (as long as seven years for the Sandy Hook massacre videos [78]), and limited by biases in the reports and by the resources available to verify them.

In this paper, we focus on *raids against YouTube videos*. We do so for the following reasons: (1) YouTube is one of the top visited sites worldwide, with more than 1 billion users and 1 billion hours of videos watched every day[1], and (2) it is plagued by aggressive behavior and extremism [52]. In fact, previous work [35] has shown that YouTube is the most heavily targeted platform by hateful and alt-right communities, and in particular 4chan's Politically Incorrect board (/pol/). Besides providing a characterization that identifies when a raid has occurred, however, previous work has not provided solutions to mitigate the problem.

In this paper, we propose a *proactive* approach towards curbing coordinated hate attacks against YouTube users. Rather than looking at attacks as they happen, or at known abusive accounts, ***we investigate whether we can automatically identify YouTube videos that are likely to be raided.*** We

*To appear at the 22nd ACM Conference on Computer-Supported Cooperative Work and Social Computing (CSCW'19).

[1] https://www.youtube.com/yt/about/press/

present a system that relies on multiple features of YouTube videos, such as title, category, thumbnail preview, as well as audio transcripts, to build a model of the characteristics of videos that are commonly raided. This also allows us to gain an understanding of *what* content attracts raids, i.e., *why* these videos are raided.

Our experiments rely on ground truth dataset of 428 raided YouTube videos obtained from our previous work [35], comparing them to 15K regular YouTube videos that were not targeted by raiders. Based on our analysis, we build classification models to assess, at upload time, whether a video is likely to be raided in the future. We rely on an *ensemble* of classifiers, each looking at a different element of the video (metadata, thumbnails, and audio transcripts), and build an ensemble detection algorithm that performs quite well, reaching AUC values of up to 94%. Overall, our work provides an important first step towards curbing raids on video sharing platforms, as we show how to detect videos targeted by coordinated hate attacks.

In summary, our paper makes the following contributions:

1. We analyze and model YouTube raids perpetrated by users of 4chan, using a ground truth dataset of 428 raided videos.
2. We build an ensemble classification model geared to determine the likelihood that a YouTube video will be raided in the future, using a number of features (video metadata, audio transcripts, thumbnails). Our system achieves an AUC of 94% when analyzing raided videos posted on 4chan with respect to all other non raided videos in our dataset.
3. We provide concrete suggestions as to how video platforms can deploy our methodology to detect raids and mitigate their impact.

2 Background & Related Work

Hate attacks on online services can happen in a number of ways. In this paper, we focus on organized attacks – *"raids"* – which are orchestrated by a community and target users on other platforms [35, 44]. In this section, we provide an overview of online raids, and describe how fringe communities organize and orchestrate them. Then, we review relevant prior work.

2.1 Anatomy of Online Raids

Unlike "typical" attacks on online services, such as denial of service [64], a raid is an attack on the community that calls a service home. The goal is not to disrupt the service itself, but rather to cause chaos and disruption to the users of the service. As such, online raids are a growing socio-technical problem. Nonetheless, it is hard to provide a precise definition of them. In the following, we offer a description of them based on previous work as well as our own observations of raids in the wild.

A prototypical raid begins with a user finding a YouTube video and posting a link to it on a third party community, e.g., 4chan's /pol/. In some cases, the original poster, or another user, might also write comments like *"you know what to do."* Shortly after, the YouTube video starts receiving a large number of negative and hateful comments. Overall, raids present a couple of key characteristics. For instance, they typically attract a large number of users, joining an effort to explicitly disrupt any productive/civil discourse that might be occurring. This is different from what would normally happen with possibly controversial videos (say, e.g., a video posted by social justice advocates), which also attract opposing points of view, though organically. The raid often generates a sudden, unexpected attack by uninvolved users.

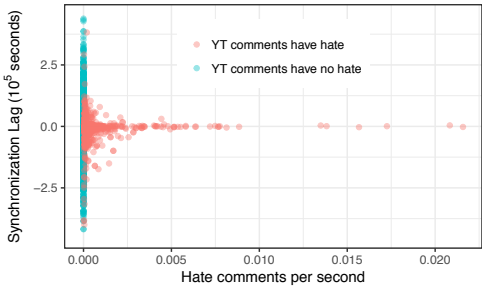


Figure 1: Distribution of videos from [35], according to the synchronization of their comments with the 4chan thread where the URL was posted and the number of hate words that appear in the comments.

Another characteristic of raids is their semi-coordinated nature. While a sudden increase in hateful comments to a video is obvious to an outside observer, what is not obvious is the fact that these comments are part of a coordinated attack. In fact, those participating in a raid may even discuss the "fruits of their labor" on the third party site that they organize on. For example, as discussed in [35], /pol/ threads serve as an aggregation point for raiders; users will post a hateful comment to the targeted YouTube video, and then brag about it on /pol/. This observation led the authors to identify videos that might have been targeted by a raid by measuring the number of "hate comments per second" (HCPS), as well as the synchronization between the comments posted on the YouTube video and those appearing on the /pol/ thread advocating for a raid.

By correlating the synchronization lag and the HCPS metric, Hine et al. [35] identified a set of videos targeted by raids during their observation period. This approach was validated by showing an increase in the overlap of YouTube accounts between videos as the synchronization lag decreases: the same accounts were more likely to be involved in the YouTube comments. In other words, it was not random YouTube users leaving hateful comments, but rather "serial" raiders almost assuredly originating from /pol/. In Figure 1, we show the distribution of videos in dataset from [35] according to synchronization lag and HCPS: observe that, the closer the lag is to zero, the higher rate of hate comments received by the video.

2.2 Related Work

YouTube. YouTube is used every day by millions of individuals to share various kinds of videos, e.g., music, lectures, gaming, video blogs, etc. [60]. Organized groups are also active on

the platform; some use it to reach and grow their network of support and organize activities, while others to propel radicalization and initiatives against other groups. Previous work has looked at the use of YouTube by LGBT users for self-disclosure [33], for anti- or pro-anorexia [57], fat stigmatization [37], sharing violent content [76], far-right propaganda [22], as well as Jihad and self-radicalization [16]. These topics often attract considerable attention and typically lead to unwanted behavior studied, e.g., in the context of swearing, abusive, hateful, or flaming comments, as well as activity against the video poster and its supporters [51, 41].

Hate on Social Platforms. Hate has driven many historical moments in the past. Even by focusing only on what happened online, researchers have studied hateful content already in the early 2000s, analyzing chat rooms behavior [30], extremist websites [28], and blogs [13]. As social networks became global communities, they also amplified discussions, creating more controversy and conflict [67]. On social networks, hate has been studied mainly by following politics and, in particular, the analysis of alt-right communities [6, 79] and populism [27].

Interactions in social communities [80] and between different groups [18] often leads to conflicts. In some cases, this may involve trolling [14, 38] or harassment [74]. Researchers have overall studied bullying and hate speech on, e.g., Twitter [9, 11] or Yahoo News and Finance [54]. Also, Salminen et al. [65] focus on a single target across platforms while Olteanu et al. [58] look at the hateful speech on Twitter and Reddit in relation to extremist violence. Mostly, social networks have reacted to this behavior through bans (e.g., on Reddit [10]) and blacklisting users (e.g., on Twitter [40]). Whereas, we aim to build a proactive, rather than reactive, system. Overall, hate speech detection systems have become a prominent area of research in the last years [23, 68, 49]. By contrast, we focus on hateful activity against YouTube videos, and in particular on studying the types of videos that are more likely to be targeted by attacks.

Cyberbullying & Aggression on YouTube. Prior work has also studied controversial YouTube videos aiming to understand what types of comments and user reaction certain categories of videos attract. For instance, Alhabash et al. [4] measure civic behavioral intention upon exposure to highly or lowly arousing videos showing cyberbullying activity, while Lange [47] studies user-posted "ranting" videos, which, although appearing to be aggressive, actually cause subtle differences in how they engage other users. Recent studies also analyze YouTube videos' comments to detect hate speech, bullying, and aggression via swearing in political videos. Kwon et al. [45, 46] investigate whether aggressive behavior (in online comments) can be contagious, observing mimicry of verbal aggression via swearing comments against Donald Trump's campaign channel. Interestingly, this aggressive emotional state can lead to contagious effects through textual mimicry. In this paper, we build on previous work on characterizing raiding behavior on YouTube, presenting a data-driven approach to identify *which* videos are likely to be the target of a raid.

Type	Source	# Videos
Raided	4chan (/pol/)	428
Non-Raided	4chan (/pol/)	789
Random	YouTube	14,444

Table 1: Overview of our datasets of YouTube videos. "Source" denotes the platform from where the link to the YouTube video was collected.

Detection. Another line of work has looked at offensive/harmful YouTube videos and how to automatically detect them. This is an orthogonal problem to ours, as we look at videos posted with a legitimate purpose, that are later victim of coordinated attacks. Sureka et al. [72] use social network analysis to identify extremist videos on YouTube, while Aggarwal et al. [2] detects violent and abusive videos, by mining the video's metadata such as linguistic features in the title and description, popularity of video, duration and category. Finally, Agarwal and Sureka [1] search for malicious and hateful videos using a topical crawler, best-first search, and shark-search for navigating nodes and links on YouTube.

Marathe and Shirsat [50] study detection techniques used for other problems, e.g., spam detection, and assess whether they could be applied to bullying detection. Also, Dadvar et al. [17] use machine learning to detect YouTube users exhibiting cyberbullying behavior. Whereas, rather than focusing on single offending users, we look at videos that are likely to receive hate and raids and their attributes from various users. Finally, Hine et al. [35], as already discussed, show that underground forums such as 4chan organize raids to platforms like Twitter, Google, and YouTube.

Remarks. In conclusion, to the best of our knowledge, this work is the first to study video properties, such as their transcript, metadata, and thumbnail, to shed light on the characteristics of the videos raided by the users of such platforms, using advanced machine and deep learning techniques to perform detection of videos targeted by raids.

3 Datasets

In this section, we introduce our three datasets used throughout the paper, as also summarized in Table 1:

1. Videos raided after being posted on /pol/, as identified by [35];
2. Videos posted on /pol/ which were *not* raided;
3. Random YouTube videos, which we use to compare raided videos against.

Raided videos posted on 4chan (ground truth). We start by collecting a ground truth dataset of raided YouTube videos. As discussed previously, fringe communities within 4chan are often responsible for organizing raids against YouTube users that promote ideas that they do not agree with. Raiders attack such videos, and being part of a group make them feel authorized to express their point of view by insulting the other users and disrupting the civil discussion on the topic. Therefore, we

obtain the dataset of YouTube links posted on 4chan over a 2.5-month period in 2016 (June to mid September) from the authors of [35]. For our purposes, we want to choose *conservative* thresholds to ensure we only select videos that we are confident were raided. Thus, based on Figure 1, we select videos with $HCPS > 10^{-4}$ and time lag less than a day, resulting in 428 videos (out of 1,217) that were raided. [See Section 2.1 for details on the Hate Comments Per Second (HPCS) metric.] We manually examined this ground truth dataset to further increase our confidence that they were indeed raided.

Non-raided videos posted on 4chan. Although many YouTube videos posted on 4chan's /pol/ are raided, obviously not all videos posted attract hateful behavior. Figure 1 shows that videos that have a high lag compared to the thread in which they are posted are unlikely to see much hateful behavior. To compare the characteristics of these videos to the raided ones, we build a second dataset with videos that were posted on 4chan but were *not* raided. We use conservative thresholds to ensure that we do not mistakenly included raided videos: to be part of this set, a video needs to have both a synchronization lag of more than one day compared to the 4chan thread it was posted in, and to have a HCPS of 0. These choices leave an unselected set of videos from the /pol/ database, making our sets cleaner and more accurate. This yields 789 non-raided videos.

Random YouTube videos. Finally, in order to draw comparisons with the ground truth of raided videos, we need to collect a set of YouTube videos that are likely not raided (we may refer to them as *not raided* from now on). We use the YouTube API and download 50 of the top videos across a variety of categories. In the end, we collected 14,444 videos, selected following the same distribution of (YouTube) categories as those linked on /pol/. We downloaded videos during the same time window as the ones posted on /pol/ to build a random set as reliable as possible, however, the dynamics of the raids may force our system to periodically update the training set.

Ethical considerations. Our research protocol received ethics approval from University College London. We also followed standard ethical practices to minimize information disclosure for all datasets, e.g., we discarded *any* personal information about the users uploading or commenting on the videos, encrypted data at rest, etc.

4 Video Processing and Analysis

We now present the methods used to analyze the characteristics of the YouTube videos in our dataset that received raids. We look at the metadata of a video, its audio transcript, as well as the thumbnail preview. We then use the insights derived from this analysis to build a classifier geared to determine whether a YouTube video is likely to receive a raid.

4.1 Metadata

In addition to the actual videos, we also collect the associated metadata, specifically: title, duration, category, description, and tags. Except for the duration, these fields are entered by the user uploading the video. Naturally, title, duration, and description often play a major role in a user's decision to watch the video as they are the first elements that they see. Also, the tags provide an intuition of a video's topics, and are actually also used by YouTube to suggest other videos—in a way, watching a suggested video might actually trigger a post on 4chan. Looking at the category for videos posted on 4chan, we anecdotally find that many of them include news, politics, and ethnic issues.

Evidence of controversial topics. We perform term frequency-inverse document frequency (*TF-IDF*) analysis on the string metadata (title, tags, and description) to extract information about the most used keywords in the different groups of videos, finding that random videos often include "Google," "music," and "love" (top 3 used words), as well as "follow" and "subscribe." By contrast, all videos posted on 4chan include politics-related words such as "Hillary" and "Trump," or indications of racial content like "black," while only raided videos have certain words like "police," "lives" (likely related to the Black Lives Matter movement), or "Alex" (referring to InfoWars' Alex Jones, a conspiracy theorist, who is well known in alt-right circles).[2]

The differences in the topics used in the metadata are extremely important: search engines are affected by the content of the metadata, especially tags; moreover YouTube suggests videos to the user based on many of these fields. Overall, we observe that random YouTube videos have few topics in common with the 4chan videos, while there are some similarities between the set of videos posted on 4chan but not raided and those that were raided.

4.2 Audio Processing

The process to extract audio from each video involve five steps. (1) We download YouTube videos in MPEG-4 format. (2) We extract the corresponding stereo audio channels using the ffmpeg tool at 44.1KHz sampling rate. (3) Both audio channels are then mixed and down-sampled to 8KHz, using the sox utility; aiming to match same acoustic conditions between the YouTube audio and the training samples employed to develop the following audio analysis modules. (4) We rely on Voice Activity Detection (VAD) to discriminate non-speech audio segments for further processing. (5) We use Automatic Speech Recognition (ASR) to perform the speech-to-text transcription. Note that previous systems were originally trained and tuned using conversational telephone speech.

Voice Activity Detection. VAD is often used as an upstream processing step intended to prevent unwanted data from entering later stages. The VAD system we use is based on [21] and uses long short-term memory (LSTM) recurrent neural networks. We train and evaluate it using call center audio, 20 hours and 1.4 hours respectively, with error rates ranging from 5% to 8%.

[2] https://en.wikipedia.org/wiki/Alex_Jones

(a) Non-raided. **(b)** Non-raided. **(c)** Raided. **(d)** Raided.

Figure 2: Sample of thumbnails from our dataset.

Automatic Speech Recognition. We use an ASR system for English from [48], trained using the Kaldi toolkit [62] and the Switchboard corpus [31], which includes around 300 hours of conversational speech. In particular, we adapt the Switchboard training recipe for nnet2 models from [62], and train two different systems. The first one makes use of GMM/HMM models and speaker adaptation techniques. It employs triphone units with a total of 5,500 states and with a final complexity of 90,000 Gaussian mixtures. The second trains a vanilla DNN, composed of 4 hidden layers with 1,024 neurons each, on top of the alignments produced from the previous GMM system.

For the language modeling, we estimate a trigram language model using MIT Language Model Toolkit with Kneser-Ney Smoothing [36]. No lattice re-scoring is performed. The pronunciation dictionary, an orthographic/phonetic mapping, is from CMUdict, an open source pronunciation dictionary.[3] The target lexicon accounts for more than 40K words. Note that neither "bad words" nor slang terms are in the original Switchboard lexicon. To evaluate the ASR performance, we use a separated subset of the same Switchboard database accounting for 5 hours of speech. The development results by the DNN based system, trained using only the Switchboard dataset, show a 13.05% Word Error Rate (WER). We finally run the DNN system on the Youtube audio dataset to generate the 1-best decoding transcription.

Evidence of controversial topics. Similar to what we did with the metadata, we also analyze the transcribed words to compare the different datasets. We observe that most YouTube videos have a lot of verbal communication. Specifically, 86% of the videos have at least 10 words spoken with the median and average video transcription containing 317 and 1,200 words respectively. We also look at whether or not some terms are more prevalent in raided YouTube videos, by averaging the *TF-IDF* vectors separately for the two classes (raided and non-raided videos), and examining the most influential terms. We find words like "black," "police," "white," "shot," "world," "gun," "war," "American," "government," and "law" in the top 20 terms in raided videos (in addition to some stop words and extremely popular terms that were excluded). Of these, the only word that appears among the top 20 in the non-raided videos is "government." The top terms for non-raided videos are different: they include words like "god," "fun," "movie," and "love."

[3] http://www.speech.cs.cmu.edu/cgi-bin/cmudict

4.3 Thumbnails

On YouTube, each video is also represented by an image thumbnail, used, e.g., in search results. Thumbnails provide viewers with a quick snapshot of the content of each video. Although users can manually select them, by default, thumbnails are automatically selected from the video and the user can choose one of three suggested options.

Using the YouTube API, we extract all available thumbnails from the videos in our dataset—specifically, using a combination of image recognition and content extraction tools (see below). Note that thumbnails are not always available to download. In a few cases, this happens because the videos is not accessible via the API when gathering the thumbnails (which was done separately from the video itself, comments, and metadata), but, mostly, because the thumbnail was not been properly uploaded and is inaccessible even though the video is still available.

Image recognition. To extract meaningful information from the thumbnails, we use deep neural networks [42, 73]. A large corpus of images can be used to train a deep neural network: each image is annotated with visible context that is used as ground truth, and the resulting network can then recognize objects appearing in the images and generate an accurate description of them.

Context extraction. For each thumbnail, we then output a description that represents the semantics involved in the image. Figure 2 shows images from four examples of different thumbnails, two in the *raided* category and two in the *non-raided* category. The following descriptions have been automatically inferred from each of the images: (a) a white plate topped with a pile of food, (b) a couple of women standing next to each other, (c) a man in a suit and tie standing in front of a TV, and (d) a woman sitting in front of a laptop computer. Note that each caption extracted not only identifies the main actor within the picture (a plate, a couple of women, or a man), but also the background activity. However, these descriptions are automatically inferred based on a model bounded by the objects in the training images-set, thus, there might be misinterpretations.

Evidence of controversial topics. We use *topic-modeling* to abstract the descriptions obtained from the images using Con-

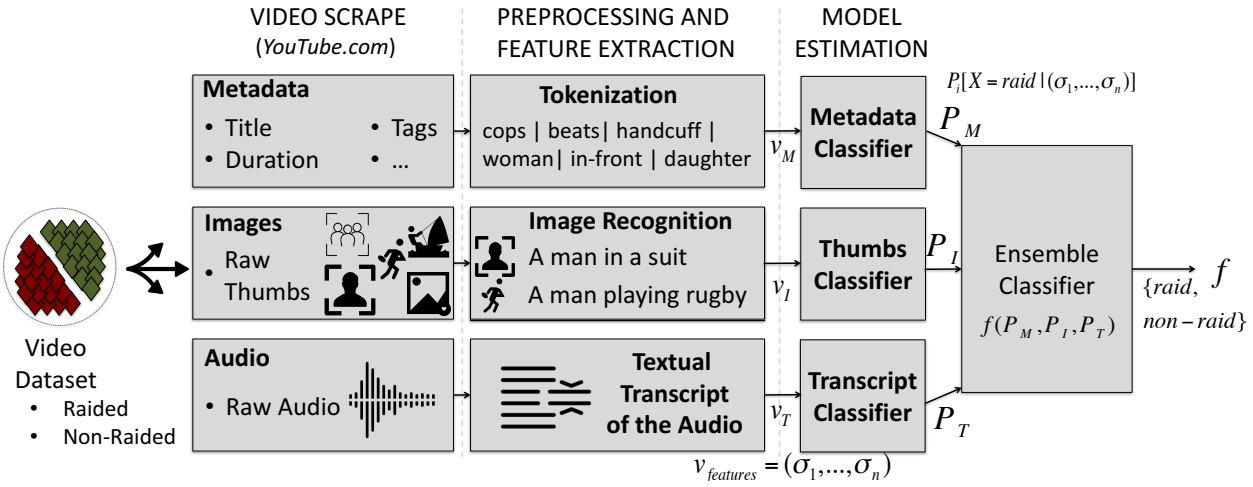


Figure 3: Architecture of our proactive detection system.

Type	Non-Raided	Raided	Diff.
Clothing	25.5%	33.4%	7.9%
Male-Gender	52.4%	59.1%	6.7%
Device	44.3%	50.7%	6.4%
Vehicle	8.9%	12.4%	3.4%
Animal	9.2%	5.8%	3.4%
Sport	22.6%	20.3%	2.2%
Color	12.5%	10.7%	1.8%
Joy	1.6%	2.8%	1.2%
Culture	1.6%	0.7%	0.9%
Food	2.4%	1.6%	0.8%
Female-Gender	9.8%	10.3%	0.5%
Nature	6.8%	6.8%	0.1%

Table 2: Topics found across videos with thumbnails.

ceptNet [69]. In particular, we extract categories related to *sports*, *joy*, *underage*, *gender*, etc. For example, some of the words in the *joy* category are "happy," "smile," "wedding," or "Christmas." Table 2 shows a summary of the prevalence of words related to any of these categories across videos in our dataset. We observe that there are a number of common topics displayed across both classes (e.g., *nature*). Likewise, female gender references are almost equal in both classes, with a slight bias towards *raid* videos. Interestingly, the case of male gender references is clearly biased towards *raided* videos, with males appearing in about 52% of the non-raided videos and in 59% of the raided ones. Reference to clothes (e.g., "tie", "dress", "jacket", "uniform") is the most distinctive category with a 7.9% difference between each class.

This indicates that there are a number of thumbnails whose context can be used to characterize videos that could potentially be raided. However, numbers also indicate that thumbnails alone might not be able to fully model the differences between the two classes. Our intuition at this point is that thumbnails can contribute towards the classification decisions, but they will not outperform other feature sets. While an exploratory analysis such as the one presented in this section can provide a general impression of the choice of models, it is hard to know upfront which model will better capture features singling out hate attacks. Extensive research work has covered the benefits of combining together different models [66]. Having different feature sets is generally the main reason for using a combination of classifiers. This is because different classification methods might perform better with a specific sub-set of features [39]. However, as pointed out in [66], there is no definitive taxonomy of combined learning; thus, an empirical comparison is paramount to determine the relative benefits and drawbacks in our domain.

5 Proactive Detection

We now introduce our approach to provide a *proactive* detection tool for videos targeted by hate attacks on online services, and on YouTube in particular. Our goal is to systematize this task, relying on a set of machine learning classifiers, each of which focuses on a different set of features extracted from online videos. Overall, we detail the set of features we use, motivated by the findings reported in the previous section.

5.1 Overview

A high-level description of our detection system is presented in Figure 3. The system is first trained using the dataset videos as explained earlier.

(1) A set of prediction models $\mathcal{C}=\{C_1,\ldots,C_i\}$ that output a probability $C_i(\sigma_1,\ldots,\sigma_n)=P_i[Y=\texttt{raid}|(\sigma_1,\ldots,\sigma_n)]$ of each video Y being raided given a feature vector $(\sigma_1,\ldots,\sigma_n)$ obtained from different elements i of the video. These models are referred to as *individual classifiers*.

(2) A weighted model $f(\mathcal{C}) = \sum w_i \cdot C_i$ that combines all predictions in $\mathcal{C}$, where each of the classifiers C_i is weighted by w_i based on the performance obtained on a validation set. This set is different from the training set (used to build the individual probabilities) and the testing set (used to measure the efficacy of the classifiers). The process of weighting the

individual predictors also serves as a way to calibrate the output of the probabilities. The final classifier will then output a decision based on its voting algorithm.

To ease presentation, we refer to the model presented in (2) as *weighted-vote*. One can simplify the model by giving equal weight to all w_i (typically $w_i = 1$) and obtaining a nominal value for C_i before voting. In other words, applying a threshold for each C_i (e.g., 0.5) and creating an equal vote among participants. We refer to this non-weighted voting system as *majority-vote*. One can further simplify the scheme by combining each individual prediction using the arithmetic mean of the output the probabilities—this is known as an *average-prediction* system.

Note that the parameters (i.e., w_i, ϵ, and the thresholds for deciding the class in each C_i) used in both *majority-vote* and *average-prediction* are fixed and do not require calibration. Thus, the validation set is not used in these two modes.

5.2 Feature Engineering

In the following, we discuss how we create the features vectors used by the different classifiers. Our system extracts features from three different sources: (1) structured attributes of the *metadata* of the video, (2) features extracted from raw *audio*, and (3) features extracted from raw *images* (thumbnails). Based on the preprocessing described earlier, we transform non-textual elements of a video (i.e., audio and images) into a text representation. Other textual elements such as the title of the video and the tags are kept as text. These textual representations are then transformed into a fixed-size input space vector of categorical features. This is done by tokenizing the input text to obtain a nominal discrete representation of the words described on it. Thus, feature vectors will have a limited number of possible values given by the bag of words representing the corpus in the training set. When extracting features from the text, we count the frequency with which a word appears in the text.

Since in large text corpus certain words–e.g., articles–can appear rather frequently without carrying meaningful information, we transform occurrences into a score based on two relative frequencies known as *term-frequency* and *inverse document-frequency* (*TF-IDF*s). Intuitively, the term frequency represents how "popular" a word is in a text (in the feature vector), and the inverse document-frequency represents how "popular" a word appears, provided that it does not appear very frequently in the corpus (the feature space). More formally, we compute as $idf(t) = log\frac{1+n_s}{1+df(s,t)} + 1$, where n_s is the total number of samples and $df(s,t)$ is the number of samples containing term t.

As for the thumbnails, after extracting the most representative descriptions per image, we remove the least informative elements and only retain entities (nouns), actions (verbs), and modifiers (adverbs and adjectives). Each element in the caption is processed to a common base to reduce inflectional forms and derived forms (known as stemming). Further, we abstract the descriptions obtained from the images using *topic-modeling* as described earlier.

In our implementation, we extract features only from the thumbnail of a video. Again, this is mainly because the thumbnails are purposely selected to and encapsulate semantically relevant context. However, we emphasize that our architecture could support the extraction of features from *every* frame in the video.

5.3 Prediction Models

We use three independent classifiers to estimate the likelihood of a video being targeted by hate attacks. These are built to operate independently, possibly when a new video is uploaded. In particular, each classifier is designed to model traits from different aspects of the video. Available decisions are later combined to provide one unified output.

We use three different classifiers, in an ensemble, because features obtained from different parts of a video are inherently incomplete, as some fields are optional and others might not be meaningful for certain video. For instance, a music video might not report a lengthy transcript, or a thumbnail might not contain distinguishable context. Since any reliable decision system should be able to deal with incomplete data, ensemble methods are well-suited to this setting. Moreover, ensembles often perform better than single classifiers overall [20]. The goal of each classifier is to extrapolate those controversial characteristics that, if present, are making the video likely to be raided.

5.3.1 Metadata and thumbnail classifiers

We build a prediction model such that $P_i(Y = \textit{raid})$ based on the features extracted from the metadata (P_M) and from the image thumbnails (P_I). The architecture of these two predictors is flexible and accepts a range of classifiers. Our current implementation supports Random Forests (RFs), Extra Randomized Trees (XTREEs), and Support Vector Machines (SVM), both radial and linear. We opt for RF as the base classifier for P_T and SVM with linear kernel for P_M. Both SVM and RF have been successfully applied to different aspects of security in the past (e.g., fraud detection [8]) and have been shown to outperform other classifiers (when compared to 180 classifiers used for real-world problems [26]).

5.3.2 Audio-transcript classifier

Before feeding the transcripts to the classifier, we remove words that have a transcription confidence $p_{trans}<0.5$, as they are likely incorrectly transcribed (including them only confuses the classifier). Note that this only removes 9.2% of transcribed words. Additionally, we remove repeated terms that are mostly exclamations such as "uh uh" or "hm hm". Finally, the transcripts contain tags for non-verbal communication such as noise, laughter, etc., which we leave in the text as they do carry predictive power. We tried classifying using traditional TF-IDF based approaches, Convolutional Networks, and Recurrent Neural Networks (RNN), ultimately opting for the latter since it yields the best performance and is quite effective at understanding sequences of words and interpreting the overall context. We also use an attention mechanism [5] that helps RNNs "focus" on word sequences that might indi-

cate potentially raided videos. We also use GloVe to embed words into 200-dimensional vectors as we have relatively few training examples.

5.3.3 Ensemble classifier

The objective of the ensemble method is to aggregate the predictions of the different base estimators. Each classifier individually models the likelihood that a video will be be targeted by hate attacks based on its set of features. The ensemble combines these decisions to make a more informed prediction. This allows for more robust predictions (in terms of confidence) and can result in a more accurate prediction. Our classifier is a stacking ensemble architecture, known to perform better than individual classifiers [77]. This architecture has been successfully used in other classification fields, for instance, credit scoring [75] and sentiment analysis [3]. We design our ensemble method to take a weighted vote of the available predictions. To compute the best-performing set of weights, we estimate a function f that takes as input each of the individual probabilities and outputs the aggregated prediction. During training this function learns from an independent testing set, and will be used during testing to weight each prediction model P_i. Formally, $f(P_M, P_I, P_T) = \{\texttt{raid}, \texttt{non-raid}\}$.

For the decision function f of our *weighted-vote* algorithm (see the overview paragraphs at the beginning of this section), we use a distribution function that models how an expected probability in the testing set is affected by individual decisions P_i in a multiple regression.

In our implementations, we use different underlying classification algorithms for estimating f. However, in the next section, we only present the results for each of the individual classifiers and two ensemble methods, namely, *weighted-vote*, and *average-prediction*. For the former, weights are fit using XTREE [29]. For the latter, we also test different settings. In particular, we try settings where one of the classifiers is given a fixed weight of $w = 0$ and we average the others.

6 Evaluation

In this section, we present the setup and the results of our experimental evaluation.

6.1 Experimental Setup

We aim to show that we can distinguish between raided and non-raided videos. However, there are several subtasks we also want to evaluate, aiming to better characterize the problem and understand how our classifiers perform.

Experiments. We start by trying to distinguish between random YouTube videos and those that are linked from /pol/. Next, we distinguish between the videos that are raided and those that are not (whether posted on /pol/ or not). Finally, we predict whether videos posted on 4chan will be raided.

More specifically, in EXPERIMENT 1, we test if our classifiers are able to distinguish between videos linked from /pol/ and a random video uploaded to YouTube, aiming to gather insight into the ability to discriminate between videos *potentially* raided vs. those that are not at risk at all. Then, EXPERIMENT 2 evaluates whether or not the classifier can distinguish between any non-raided video (i.e., regardless of whether it is a random YouTube video or one posted on 4chan) and videos that will be raided. Finally, in EXPERIMENT 3, we focus on videos posted on 4chan, and determine which are going to be raided and which are not; this ensures that we can not only predict whether a video was posted on 4chan, but whether or not the video will be raided.

Train, Test, and Validate Splits. We split our datasets into three chunks: two for training and tuning parameters of the ensemble (training and validation) and one for testing, and report performance metrics on the latter. As we are dealing with highly unbalanced classes (there are multiple orders of magnitude more videos posted to YouTube than those posted to 4chan, let alone those that are raided), we balance the training and validation sets to model both classes properly, but leave the test set unbalanced. We have tried to train and tune the classifiers using unbalanced sets, but preliminary results have shown that this setup does not allow an accurate classification. This setup underperforms by at least 0.12 AUC in all the different experiments compared to the results we had by training and tuning on balanced sets. The test set, however, remains unbalanced to more realistically model the ratio of raided videos against non raided videos in the wild.

The total number of videos in each split is proportionally sampled depending on the less populated class, assigning splits of 60%, 20%, and 20% to the training, validation, and test sets. The more populated class uses the same amount of samples for training and validation, while it will have all the remaining samples in the test set. This procedure is repeated ten times and the results are averaged over the ten different rounds. We decided to use random sampling for the training set, rather than a stratified sampling based on the categories of the videos as we believe this allows us to present results from a worst-case scenario: training videos may not fully represent test videos and, as consequence, our classifiers may perform slightly worse. Table 3 summarizes all settings in our experiments, along with the number of samples used.

Evaluation Metrics. We evaluate our system using precision, recall, and F1-measure. Overall, these metrics are a good summary of the performance of an classifier in terms of True Negatives (TN), False Negatives (FN), False Positives (FP), and True Positives (TP); however, they are not ideal for comparing results across different experiments. Therefore, we also plot the Area Under Curve (AUC), which reports the TP-rate (*recall*) against the FP-rate (*1 - recall*).

6.2 Experimental Results

We now report the results of our experimental evaluations, as per the settings introduced above. To ease presentation, we only report metrics for the individual classifiers as well as two ensemble methods: *weighted-vote* and *average-prediction*. We do not report results for other ensemble classifiers (*simple-voting* and the other underlying algorithms for estimating the

ID	Description	Training	Validation	Test
EXPERIMENT 1	Random YouTube vs. all on 4chan	731+731	243+243	13,470+244
EXPERIMENT 2	All non-raided vs. raided on 4chan	258+258	85+85	14,890+85
EXPERIMENT 3	Non-raided on 4chan vs. raided on 4chan	258+258	85+85	446+85

Table 3: Number of samples used in our experiments. The sets are balanced as there is the same amount of samples per each class (*class 1 samples+class 2 samples*) in training and validation, while they are unbalanced in the test set.

	EXPERIMENT 1				EXPERIMENT 2				EXPERIMENT 3			
Classifier	**PRE**	**REC**	**F1**	**AUC**	**PRE**	**REC**	**F1**	**AUC**	**PRE**	**REC**	**F1**	**AUC**
transcripts	0.05	0.60	0.10	0.79	0.03	0.56	0.06	0.79	0.32	0.58	0.40	0.73
metadata	0.13	0.89	0.23	**0.96**	0.03	0.85	0.06	**0.94**	0.32	**0.71**	0.44	0.79
thumbnails	0.02	0.64	0.05	0.62	0.01	0.66	0.02	0.61	0.18	0.55	0.27	0.56
weighted-vote ensemble	0.12	**0.91**	0.21	**0.96**	0.03	**0.88**	0.05	**0.94**	0.34	0.69	0.45	**0.80**
average-prediction ensemble	**0.15**	0.85	**0.26**	**0.96**	**0.04**	0.82	**0.07**	**0.94**	**0.35**	0.69	**0.46**	**0.80**

Table 4: Results for EXPERIMENT 1, EXPERIMENT 2, and EXPERIMENT 3. PRE stands for precision, and REC for recall. The ensemble classifiers have different inputs: the weighted-vote classifier receives inputs from all three the individual ones, while the average-prediction does not receive the thumbnail classifier input.

weights), since they under-perform in our experiments. For *weighted-vote*, weights are fit using XTREE [29], as described earlier. Also note that, for *average-prediction*, we find that the thumbnails classifier tends to disagree with the metadata and the transcripts classifiers combined. Therefore, for *average-prediction*, we fix a weight of $w = 0$ for the thumbnails classifier (i.e., $w_{thumbnail} = 0$).

Experiment 1. In this experiment we study whether we can predict that a video is linked from 4chan. Results are reported in Table 4. Since we are dealing with a rather unbalanced validation set (in favor of the negative class), it is not surprising that precision drops to values close to 0 even though we maintain high recall.

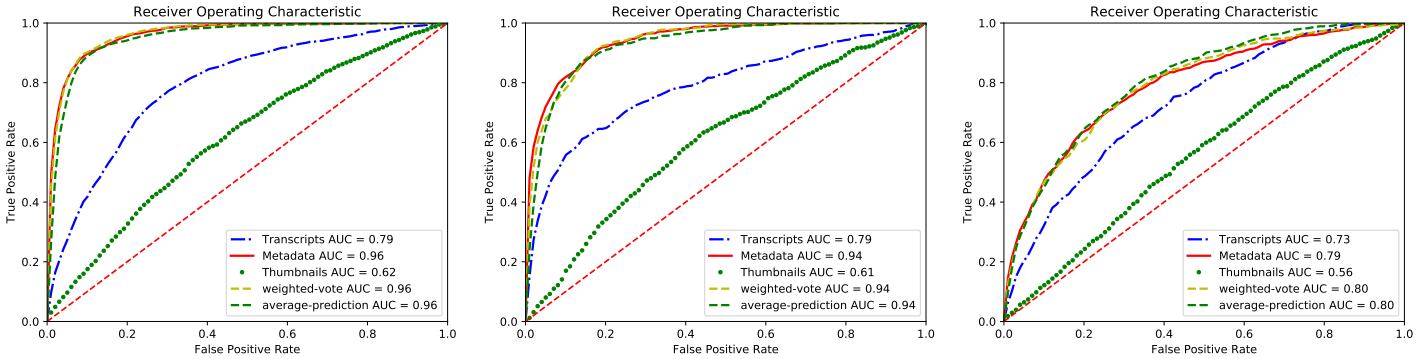
Looking at the results obtained by the individual classifiers, the weighted-vote ensemble classifier matched the best recall from the metadata individual classifier (0.91). The model relied almost entirely on the metadata classifier (weight equal to 0.987) while assigning very low weights to transcripts (0.007) and thumbnails (0.006). The best AUC is the same between the metadata classifier and the two ensemble classifiers (0.96).

In Figure 4a, we also plot the ROC curve for all five classifiers. The individual AUC scores are 0.79, 0.96, 0.62 for the transcripts, metadata, and thumbnails, respectively, while the two ensembles (*weighted-vote* and *average-prediction*) score 0.96. The *weighted-vote* ensemble has the highest AUC throughout most of the x-axis, although, the ROC curve essentially overlaps with that of the metadata classifier. The two ensembles have different strengths: the *weighted-vote* has the highest recall and AUC values, but the *average-prediction* (with $w_{thumbnail} = 0$) has highest precision, and F1-measure.

Experiment 2. In Figure 4b, we report the AUC when classifying raided and non-raided videos—regardless of whether the latter are random YouTube videos or non-raided ones posted on 4chan. Unlike EXPERIMENT 1, among the individual classifiers, the best performance is achieved by the audio-transcript classifier, except for recall. This setting also yields high recall (0.88) when combining all classifiers into the *weighted-vote* ensemble. As in EXPERIMENT 1, the *weighted-vote* ensemble presents the highest recall and AUC, but the *average-prediction* has higher precision, and F1-measure. Once again, the model relied almost entirely on the metadata classifier (weight equal to 0.984) while assigning very low weights to transcripts (0.008) and thumbnails (0.008). Figure 4b shows a similar situation as in the previous experiment: the ROC curve for the metadata classifier is really close to or overlapping with the ones for the two ensemble. AUC equals to 0.61 for thumbnails, 0.79 for transcripts, and 0.94 for metadata. Whereas, the two ensemble classifiers achieve 0.94 AUC as the metadata individual classifier

Experiment 3. Finally, we evaluate how well our models discriminate between raided videos posted to 4chan and non-raided videos also posted to 4chan. Our results confirm that this is indeed the most challenging task. Intuitively, these videos are much more similar to each other than those found randomly on YouTube as /pol/ is interested in a particular type of content.

This setting shows a clear case for the ensemble classification yielding appreciably better performance. Overall, the individual classifiers, i.e., transcripts, metadata, and thumbnails, reach AUCs of 0.73, 0.79, and 0.56, respectively, whereas, both the ensemble classifiers reach 0.80. Nevertheless, the ROC curve in Figure 4c shows how the *weighted-vote* ensemble is sometimes penalized by the weakest performing classifier (i.e., thumbnails classifier). This is apparent by comparing the *weighted-vote* and *average-prediction* (recall that $w_{thumbnails} = 0$ in the latter). The weights assigned by the model confirm this: the metadata classifier's weight is 0.868, while the weight for transcripts is 0.034 and the one for thumbnails is 0.098. As expected, the metadata has a very important weight (although not as high as in the previous experiments), surprisingly, the thumbnails have a higher weight than the transcripts. This means that the transcripts classifier tend to agree with the metadata one in most of the cases they correctly flag the videos, while the thumbnails classifier identifies



(a) Experiment 1 **(b)** Experiment 2 **(c)** Experiment 3

Figure 4: ROC curves for each experiment. AUC values for Thumbnails, Transcripts, Metadata, and Ensemble classifiers, XTREE, and Average probabilities. This metric provides a fair comparison across experiments.

correctly some videos when the other classifiers fail.

6.3 Choosing an Ensemble

The goal of the our system is to make the best final "decision" possible given the choices made by the individual classifiers. In absolute terms, the *weighted-vote* (with XTREE as baseline estimator) yields the best performance in all three experiments in terms of recall (and overall AUC). In particular, it outperforms the *average-prediction* ensemble in two of the tasks: modeling videos from /pol/ (EXPERIMENT 1), and detecting raided videos in general (EXPERIMENT 2). When restricting our attention to the detection of raids of videos posted on 4chan (EXPERIMENT 3), both ensemble methods are comparable in terms of recall. However, when looking at precision, we find that *average-prediction* outperforms *weighted-vote*. The trade-off between having good precision and good recall will have an important impact on the amount of vetting work required by providers deploying our system, as we discuss in more detail in Section 7.

In the following, we provide an explanation as to why the two ensemble classifiers report similar results in some experiments and different ones in others. When using a base estimator to fit the best weights for the individual classifiers, we observe a bias towards the decisions made by the metadata classifier. This is expected, as this classifier is the one that performs best among the individual classifiers (and substantially so in both EXPERIMENT 1 and EXPERIMENT 2). On the contrary, the thumbnails classifier performs worst, except for recall in EXPERIMENT 2.

Note that our data include videos with partially-available features. When this happens, the ensemble classifier is forced to make a decision based on the other inputs. It is the thumbnails case, which are not always available. This is why we evaluated the *average-prediction* ensemble method forcing a weight $w_{thumbnails} = 0$. In this setting, the *weighted-vote* method with XTREE provided similar results, since XTREE initially assigned a low weight (although not exactly 0) to the thumbnails.

Overall, with the *average-prediction* method, precision, and F1-measure are always better than for the XTREE ensemble classifiers. This means that this configuration reduces the number of false positives and, as a consequence, is slightly more accurate. Therefore, when the individual classifiers have similar performance, the ensemble is better than the best options among the single classifiers.

7 Discussion

Our experiments show that we can model YouTube videos and predict those likely to be raided by off-platform communities. In other words, our results indicate that it is possible to develop automated techniques to mitigate, and possibly prevent, the socio-technical problem of online attacks and harassment. What still needs to be ironed out is how our techniques could be integrated and deployed by online services like YouTube. Although devising a path to adoption is beyond the scope of this paper, we discuss this aspect later in this section.

Data collection. The data collection involved in this work presented a variety of challenges. For example, how can we be sure that the HCPS metric is effective in identifying hateful comments? How do we know that non raided videos on /pol/ or random videos are not False Negative cases? How can we ensure that videos are actually random to a reasonable extent? Balancing these questions while aiming to support large-scale, automated analysis thus played a large role in our design decisions.

The use of HCPS and time lag metrics from [35] allowed us to evaluate the system on ground-truth videos that were already discovered by previous work. Unfortunately, as expected, some of these videos were quite popular, with an extremely large number of comments making manual checking not viable. Conversely, using automatically quantified metrics may result in assigning the wrong class to a video (e.g., HCPS may not identify hateful comments in a specific video). To mitigate this issue, we took a conservative approach, constraining our dataset to videos that, e.g., had a minimum HCPS

and a time lag of less than a day. While this does leave questions about "borderline" cases, it also results in less ambiguity surrounding our ground truth.

Also note that the random videos dataset was collected by category, aiming to reproduce the same distribution of categories as our ground truth. As YouTube does not provide detailed statistics on the overall distribution of videos, we opted to shape our negative class following the worst possible case: an exact copy of the /pol/ distribution. That is, our negative class comprises videos that (at least in terms of category) /pol/ might find "interesting." Further, collecting the negative dataset via API queries removes additional concerns about human decisions influencing its composition. A limitation of this approach is that, by relying on /pol/ videos and the HCPS data from previous work, we were not able to apply the algorithm to the random videos. However, such limitation may only have negative effects on our results, meaning that our experiments yielded a lower bound on the system efficiency.

Evaluation. Our evaluation shows that we can reliably distinguish videos that are likely to be raided from regular YouTube videos. This means that YouTube could run our system at *upload time*, determining the likelihood that a video will be raided at some point in the future. The platform could then adopt mitigation strategies for "risky" videos, e.g., by manually moderating comments. This is a practice already employed by YouTube [61], however, at the moment, the efficacy of this moderation has been questioned due to the sheer volume of content, in addition to YouTube's focus on removing *inappropriate* videos rather than protecting users against raids.

Using our system, we estimate that only 16% of videos would require any action—an estimation based on EXPERIMENT 2. While this might appear to be a very large number, it is still less than having to monitor all videos. Moreover, additional automated tools could be deployed to check whether a raid is actually occurring before being passed along for human review. Furthermore, YouTube has pledged to hire 10,000 new workers to monitor content about a year ago [25], but deploying our system could reduce the need for this human workforce, or at worst, allow them to focus on higher impact content.

In addition, EXPERIMENT 3 demonstrates that, when provided with videos linked from fringe communities such as /pol/, our system can successfully identify videos likely to be raided with reasonable accuracy. This is a much more difficult task, and thus the correct detections are less than before, since videos are very often being posted to /pol/ without the actual intent of having raiders show up in the first place. Furthermore, the number of videos posted on /pol/ is much smaller than those uploaded to YouTube overall—for instance, the /pol/ dataset from [35] contains links to 93k YouTube videos posted over 3 months. Among these, we only selected those that had clear evidence of raids by restricting the thresholds of HCPS and time lag (see Section 2.1), also discarding videos which could have been raided from the non-raided group. This choice is extremely conservative (yielding 428 videos), aiming to have a reliable ground truth on which to evaluate the system. Although we could have relaxed our thresholds and obtained better results overall, the applicability to real-world use cases would likely have been affected, as our ground truth dataset would have included videos that had controversial or combative statements, but were not actually raided.

Adoption. Our system is really geared to identify videos that are *at risk* of being raided, thus, YouTube could use it as a *filter* — flagging videos that are risky and/or might require particular attention. We emphasize that an automated system such as the one presented in this paper inherits classical machine-learning limitations, e.g., dealing with misclassifications. However, the threshold for what is considered actionable can be tuned, since in the end the classifiers output a probability as opposed to a purely binary classification. In other words, a deployed system could be adjusted to focus on minimizing false negatives (i.e., videos that will be raided but are misclassified by the system) while still maintaining high recall. Tuning the model to balance the overall number of flagged videos as well as ensuring that they are indeed at high risk can help reduce the impact of false positives. Given that our datasets are extremely unbalanced, high precision, on the other hand, is not a top priority. As mentioned above, the system would flag videos likely to be raided, thus, helping to tackle the problem of aggression by reducing the videos that need to be monitored as high risk ones. Overall, we envision the system not as an end-all be-all solution, but as an early warning system, enhancing other mechanisms the platform decides to put in place At minimum, e.g., allowing them to focus human efforts (content moderators etc.) their efforts where abusers are most likely to attack. While this would certainly reduce the amount of human labor involved in dealing with raids, it could also be used to focus more expensive algorithmic systems on high risk videos. E.g., by introducing a post-prediction system leveraging the HCPS metric discussed earlier (c.f. Section 2.1) as well as other existing mechanisms to flag offending comments [12].

Limitations. As mentioned previously, our data collection attempted to minimize the mislabeling of videos linked on /pol/. This approach allowed us to be reasonably confident that the raided videos dataset contained only videos attacked by the /pol/ community. However, these constraints limited the size of our dataset (out of the more than $5,000$ videos linked on /pol/ we considered only 428 videos as having been raided). This relatively small dataset was a limiting factor in the performance of our classifiers.

Moreover, we assume that negative class dataset does not contain mislabeled samples. The assumption may not hold up 100% of the time since labeling errors occur with both manual and automatic checks. That said, while mislabeling in our negative class dataset might affect classifier performance to some degree, the overall validity of our system is not affected.

Also note that /pol/, though a very good example of a tight-knit community used to coordinate and disrupt other social groups, is not the only community responsible for performing raids against YouTube videos. Other Web communities, e.g,.

Reddit [63] or Kiwi Farms[4] also regularly take part in raiding activity. The same techniques presented here, however, can be used to detect raids from other communities.

Raids and communities. Finally, it might be tempting to dismiss the relatively low occurrence of raids, vis-à-vis the number of YouTube videos posted every day, as being a niche problem. On the contrary, harassment and bullying on YouTube are widely recognized as a serious issue by authorities on the matter [55], and news reports are filled with ghastly stories [59] and advice on how to deal with hateful and harassing YouTube comments in particular [19, 32].

Although we are not aware of any suicide cases directly linked to YouTube raids, victims have indeed been active on YouTube [56] and thus raids pose very serious safety risks. Overall, even if the majority of content on YouTube (and other social media platforms) tends to be "safe," we should not discard the outsized effects that this negative behavior has. From a purely pragmatic point of view, advertisers providing the primary income stream for sites like YouTube have been rethinking their reliance on social media in light of the recent surge in anti-social behavior [70]. From a societal point of view, raiding behavior is a pressing concern; it is a direct threat to free speech and civil discourse, and causes emotional distress that can lead to dire consequences. The efforts of the research community have enabled the long tail of the Web to succeed, building technologies that democratized information and shrunk the world. Thus, while raids on YouTube videos do occur in the long tail, we argue that dismissing them as being too rare is an abdication of our social responsibility.

8 Conclusion

This paper presented a supervised learning based approach to automatically determine whether a YouTube video is likely to be "raided," i.e., receive a sudden spike in hateful comments as a result of an orchestrated effort coordinated from another platform. Our experimental results showed that even single-input classifiers that use metadata, thumbnails, or audio transcripts can be effective, and that an ensemble of classifiers can reach high detection performance, thus providing a deployable early-warning system.

Overall, our work represents an important first step toward providing video platforms like YouTube with proactive systems geared to detect and mitigate coordinated hate attacks. We discussed potential deployment strategies that could be taken by YouTube (or other providers), i.e., running our tool on every video at upload time and/or monitoring fringe communities such as 4chan to screen videos that are linked to on those platforms.

Note that the classifiers presented in this paper are not meant to provide a mechanism for *censoring* content or users, nor to identify users possibly involved in raids. Rather, we aim to identify content that is at risk of attack; once identified, *proactive* solutions to protect against raiders can be taken by the service providers. While the specifics are beyond the purpose of this paper, we believe that there are actions that can be taken that protect freedom of expression while also preserving civil discourse. For example, temporarily disabling or rate limiting comments, requiring new comments to be approved before going live, or simply notifying the poster that a raid might be coming could serve to balance protection vs. expression.

As part of future work, we plan to use rank aggregation techniques as ensemble, as well as deep-learning methods to fuse audio, video, and metadata into a single classifier. This design follows a different approach with respect to the current one. It needs more data (and as consequence more time and computational resources for training) than the amount of raided videos we currently have, but, as it can manage the relations among the different features types, it is likely to have better performances once trained properly. We also plan to look into raids from other communities, such as Reddit, Gab.ai, and Kiwi Farms.

Acknowledgments. This project has received funding from the European Union's Horizon 2020 Research and Innovation program under the Marie Skłodowska-Curie ENCASE project (GA No. 691025). Enrico Mariconti was also supported by the EPSRC under grant 1490017.

[4] https://kiwifarms.net/

References

[1] S. Agarwal and A. Sureka. A Focused Crawler for Mining Hate and Extremism Promoting Videos on YouTube. In *ACM Hypertext*, 2014.

[2] N. Aggarwal, S. Agrawal, and A. Sureka. Mining YouTube metadata for detecting privacy invading harassment and misdemeanor videos. In *PST*, 2014.

[3] S. Al-Azani and E.-S. M. El-Alfy. Using word embedding and ensemble learning for highly imbalanced data sentiment analysis in short arabic text. *Procedia Computer Science*, 109, 2017.

[4] S. Alhabash, J. hwan Baek, C. Cunningham, and A. Hagerstrom. To comment or not to comment?: How virality, arousal level, and commenting behavior on YouTube videos affect civic behavioral intentions. *Computers in Human Behavior*, 51, 2015.

[5] D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. *arXiv:1409.0473*, 2014.

[6] A. Ben-David and A. Matamoros-Fernández. Hate speech and covert discrimination on social media: Monitoring the Facebook pages of extreme-right political parties in Spain. *International Journal of Communication*, 10, 2016.

[7] F. Benevenuto, G. Magno, T. Rodrigues, and V. Almeida. Detecting spammers on twitter. In *CEAS*, 2010.

[8] S. Bhattacharyya, S. Jha, K. Tharakunnel, and J. C. Westland. Data mining for credit card fraud: A comparative study. *Decision Support Systems*, 50(3), 2011.

[9] P. Burnap and M. L. Williams. Us and them: identifying cyber hate on Twitter across multiple protected characteristics. *EPJ Data Science*, 5, 2016.

[10] E. Chandrasekharan, U. Pavalanathan, A. Srinivasan, A. Glynn, J. Eisenstein, and E. Gilbert. You can't stay here: The efficacy of reddit's 2015 ban examined through hate speech. *Proceed-*

ings of the ACM on Human-Computer Interaction, 1(CSCW), 2017.

[11] D. Chatzakou, N. Kourtellis, J. Blackburn, E. De Cristofaro, G. Stringhini, and A. Vakali. Hate is not Binary: Studying Abusive Behavior of #GamerGate on Twitter. In *ACM Hypertext*, 2017.

[12] D. Chatzakou, N. Kourtellis, J. Blackburn, E. De Cristofaro, G. Stringhini, and A. Vakali. Mean Birds: Detecting Aggression and Bullying on Twitter. In *International ACM Web Science Conference*, 2017.

[13] M. Chau and J. Xu. Mining communities and their relationships in blogs: A study of online hate groups. *International Journal of Human-Computer Studies*, 65(1), 2007.

[14] J. Cheng, M. Bernstein, C. Danescu-Niculescu-Mizil, and J. Leskovec. Anyone can become a troll: Causes of trolling behavior in online discussions. In *Proceedings of the ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW)*, 2017.

[15] S. Chess and A. Shaw. A conspiracy of fishes, or, how we learned to stop worrying about #GamerGate and embrace hegemonic masculinity. *Journal of Broadcasting & Electronic Media*, 2015.

[16] M. Conway and L. McInerney. *Jihadi Video and Auto-radicalisation: Evidence from an Exploratory YouTube Study*. Springer Berlin Heidelberg, 2008.

[17] M. Dadvar, R. Trieschnigg, and F. de Jong. Experts and Machines against Bullies: A Hybrid Approach to Detect Cyberbullies. In *Canadian AI*, 2014.

[18] S. Datta, C. Phelan, and E. Adar. Identifying misaligned intergroup links and communities. *Proceedings of the ACM on Human-Computer Interaction*, 1(CSCW), 2017.

[19] E. Denton. YouTuber Jaclyn Hill Reveals She's "Scared" of Her Channel Because the Comments Are So Mean. https://bit.ly/2HXSTa4, 2015.

[20] T. G. Dietterich. Ensemble Methods in Machine Learning. In *First International Workshop on Multiple Classifier Systems*, 2000.

[21] D. B. Eichenberger. Speech activity detection: Application-specific tuning and context-based neural approaches. Bachelor thesis, Universitat Politècnica de Catalunya, July 2016.

[22] M. Ekman. The dark side of online activism: Swedish right-wing extremist video activism on YouTube. *MedieKultur: Journal of media and communication research*, 30(56), 2014.

[23] M. ElSherief, V. Kulkarni, D. Nguyen, W. Y. Wang, and E. Belding. Hate lingo: A target-based linguistic analysis of hate speech in social media. In *International Conference on Web and Social Media (ICWSM)*, 2018.

[24] M. ElSherief, S. Nilizadeh, D. Nguyen, G. Vigna, and E. Belding. Peer to Peer Hate: Hate Speech Instigators and Their Targets. In *International Conference on Web and Social Media (ICWSM)*, 2018.

[25] B. Feldman. Can 10,000 Moderators Save YouTube? http://nymag.com/selectall/2017/12/can-10-000-moderators-save-youtube.html, 2017.

[26] M. Fernández-Delgado, E. Cernadas, S. Barro, and D. Amorim. Do we need hundreds of classifiers to solve real world classification problems? *JMLR*, 15(1), 2014.

[27] P. Gerbaudo. Social media and populism: an elective affinity? *Media, Culture & Society*, 40(5), 2018.

[28] P. B. Gerstenfeld, D. R. Grant, and C.-P. Chiang. Hate online: A content analysis of extremist Internet sites. *Analyses of social issues and public policy*, 3(1), 2003.

[29] P. Geurts, D. Ernst, and L. Wehenkel. Extremely randomized trees. *Machine learning*, 63(1), 2006.

[30] J. Glaser, J. Dixit, and D. P. Green. Studying hate crime with the internet: What makes racists advocate racial violence? *Journal of Social Issues*, 58(1), 2002.

[31] J. J. Godfrey, E. C. Holliman, and J. McDaniel. SWITCHBOARD: Telephone Speech Corpus for Research and Development. In *ICASSP*, 1992.

[32] P. Gomez. YouTube and Instagram Stars Explain How to Protect Your Kids from Online Bullying. http://people.com/social-media-stars/protect-kids-online-bullying-youtube-instagram/, 2017.

[33] M. Green, A. Bobrowicz, and C. S. Ang. The lesbian, gay, bisexual and transgender community online: discussions of bullying and self-disclosure in YouTube videos. *Behaviour & Information Technology*, 2015.

[34] D. W. Grigg. Cyber-aggression: Definition and concept of cyberbullying. *Australian Journal of Guidance and Counselling*, 20(2), 2010.

[35] G. E. Hine, J. Onaolapo, E. De Cristofaro, N. Kourtellis, I. Leontiadis, R. Samaras, G. Stringhini, and J. Blackburn. Kek, Cucks, and God Emperor Trump: A Measurement Study of 4chan's Politically Incorrect Forum and Its Effects on the Web. In *AAAI ICWSM*, 2017.

[36] B.-J. P. Hsu and J. R. Glass. Iterative language model estimation: efficient data structure & algorithms. In *Interspeech*, 2008.

[37] M. Hussin, S. Frazier, and J. K. Thompson. Fat stigmatization on YouTube: A content analysis. *Body Image*, 8(1), 2011.

[38] A. Israni, S. Erete, and C. L. Smith. Snitches, Trolls, and Social Norms: Unpacking Perceptions of Social Media Use for Crime Prevention. In *Proceedings of the ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW)*, 2017.

[39] A. K. Jain, R. P. W. Duin, and J. Mao. Statistical pattern recognition: A review. *IEEE Transactions on pattern analysis and machine intelligence*, 22(1), 2000.

[40] S. Jhaver, S. Ghoshal, A. Bruckman, and E. Gilbert. Online harassment and content moderation: The case of blocklists. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 25(2), 2018.

[41] L. Jönson. Flaming motivation in YouTube users as a function of the traits Disinhibition seeking, Assertiveness and Anxiety? Technical report, University of Twente, 2013.

[42] A. Karpathy and L. Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

[43] I. Kayes, N. Kourtellis, D. Quercia, A. Iamnitchi, and F. Bonchi. The Social World of Content Abusers in Community Question Answering. In *WWW*, 2015.

[44] S. Kumar, W. L. Hamilton, J. Leskovec, and D. Jurafsky. Community Interaction and Conflict on the Web. In *The Web Conference (WWW)*, 2018.

[45] K. Kwon and A. Gruzd. Is Aggression Contagious Online? A Case of Swearing on Donald Trump's Campaign Videos on YouTube. In *Hawaii International Conference on System Sciences*, 2017.

[46] K. H. Kwon and A. Gruzd. Is offensive commenting contagious online? Examining public vs interpersonal swearing in response to Donald Trump's YouTube campaign videos. *Internet Research*, 2017.

[47] P. G. Lange. Commenting on YouTube rants: Perceptions of inappropriateness or civic engagement? *Journal of Pragmatics*,

2014.

[48] J. Luque, C. Segura, A. Sánchez, M. Umbert, and L. A. Galindo. The Role of Linguistic and Prosodic Cues on the Prediction of Self-Reported Satisfaction in Contact Centre Phone Calls. In *Proc. Interspeech 2017*, 2017.

[49] S. K. Maity, A. Chakraborty, P. Goyal, and A. Mukherjee. Opinion Conflicts: An Effective Route to Detect Incivility in Twitter. *Proc. ACM Hum.-Comput. Interact.*, 2(CSCW), 2018.

[50] S. Marathe and K. P. Shirsat. Approaches for Mining YouTube Videos Metadata in Cyber bullying Detection. *International Journal of Engineering Research & Technology*, 4, 2015.

[51] P. J. Moor, A. Heuvelman, and R. Verleur. Flaming on YouTube. *Computers in Human Behavior*, 26(6), 2010.

[52] B. Moser. How YouTube Became the Worldwide Leader in White Supremacy. https://newrepublic.com/article/144141/youtube-became-worldwide-leader-white-supremacy, 2017.

[53] S. Nilizadeh, F. Labrèche, A. Sedighian, A. Zand, J. Fernandez, C. Kruegel, G. Stringhini, and G. Vigna. POISED: Spotting Twitter Spam Off the Beaten Paths. In *ACM SIGSAC Conference on Computer and Communications Security (CCS)*, 2017.

[54] C. Nobata, J. Tetreault, A. Thomas, Y. Mehdad, and Y. Chang. Abusive Language Detection in Online User Content. In *WWW*, 2016.

[55] Nobullying.com. Youtube Bullying. https://nobullying.com/youtube-bullying/, 2018.

[56] A. O'Connor. Suicide Draws Attention to Gay Bullying. https://well.blogs.nytimes.com/2011/09/21/suicide-of-gay-teenager-who-urged-hope/, 2011.

[57] A. Oksanen, D. Garcia, A. Sirola, M. Näsi, M. Kaakinen, T. Keipi, and P. Räsänen. Pro-Anorexia and Anti-Pro-Anorexia Videos on YouTube: Sentiment Analysis of User Responses. *Journal of Medical Internet Research*, 17, 2015.

[58] A. Olteanu, C. Castillo, J. Boy, and K. R. Varshney. The effect of extremist violence on hateful speech online. In *International Conference on Web and Social Media (ICWSM)*, 2018.

[59] M. Oppenheim. Jessi Slaughter on becoming a meme and falling victim to trolls after infamous YouTube video. https://ind.pn/2IjeRnB, 2016.

[60] J. Y. Park, J. Jang, A. Jaimes, C.-W. Chung, and S.-H. Myaeng. Exploring the User-generated Content (UGC) Uploading Behavior on Youtube. In *WWW Companion*, 2014.

[61] S. Perez. YouTube promises to increase content moderation and other enforcement staff to 10K in 2018. https://goo.gl/2te7HV, 2018.

[62] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely. The Kaldi Speech Recognition Toolkit. In *IEEE Workshop on Automatic Speech Recognition and Understanding*, 2011.

[63] A. Romano. Reddit just banned one of its most toxic forums. But it won't touch The_Donald. https://www.vox.com/culture/2017/11/13/16624688/reddit-bans-incels-the-donald-controversy, 2017.

[64] C. Rossow. Amplification Hell: Revisiting Network Protocols for DDoS Abuse. In *Network and Distributed Systems Security Symposium (NDSS)*, 2014.

[65] J. Salminen, H. Almerekhi, M. Milenković, S.-g. Jung, J. An, H. Kwak, and B. J. Jansen. Anatomy of online hate: developing a taxonomy and machine learning models for identifying and classifying hate in online news media. In *International Conference on Web and Social Media (ICWSM)*, 2018.

[66] M. Sewell. Ensemble learning. *RN*, 11(02), 2008.

[67] P. Sobkowicz and A. Sobkowicz. Dynamics of hate based Internet user networks. *The European Physical Journal B*, 73(4), 2010.

[68] D. Soni and V. K. Singh. See No Evil, Hear No Evil: Audio-Visual-Textual Cyberbullying Detection. *Proc. ACM Hum.-Comput. Interact.*, 2(CSCW), 2018.

[69] R. Speer, J. Chin, and C. Havasi. ConceptNet 5.5: An Open Multilingual Graph of General Knowledge. In *AAAI*, 2017.

[70] N. Statt. YouTube is facing a full-scale advertising boycott over hate speech. https://www.theverge.com/2017/3/24/15053990/google-youtube-advertising-boycott-hate-speech, 2017.

[71] G. Stringhini, P. Mourlanne, G. Jacob, M. Egele, C. Kruegel, and G. Vigna. Evilcohort: detecting communities of malicious accounts on online services. In *USENIX Security Symposium*, 2015.

[72] A. Sureka, P. Kumaraguru, A. Goyal, and S. Chhabra. Mining YouTube to Discover Extremist Videos, Users and Hidden Communities. In *AIRS*, 2010.

[73] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: Lessons learned from the 2015 MSCOCO image captioning challenge. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 2016.

[74] J. Vitak, K. Chadha, L. Steiner, and Z. Ashktorab. Identifying Women's Experiences With and Strategies for Mitigating Negative Effects of Online Harassment. In *Proceedings of the ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW)*, 2017.

[75] G. Wang, J. Hao, J. Ma, and H. Jiang. A comparative assessment of ensemble learning for credit scoring. *Expert systems with applications*, 38(1), 2011.

[76] A. Weaver, A. Zelenkauskaite, and L. Samson. The (Non)Violent World of Youtube: Content Trends in Web Video. *Journal of Communication*, 62(6), 2012.

[77] D. H. Wolpert. Stacked generalization. *Neural networks*, 5(2), 1992.

[78] K. Yurieff. Google's CEO knows YouTube must do better at policing hate. https://edition.cnn.com/2019/06/17/tech/youtube-lgbt-google-ceo-sundar-pichai/index.html, 2019.

[79] S. Zannettou, B. Bradlyn, E. De Cristofaro, H. Kwak, M. Sirivianos, G. Stringini, and J. Blackburn. What is gab: A bastion of free speech or an alt-right echo chamber. In *CyberSafety*, 2018.

[80] J. Zhang, C. Danescu-Niculescu-Mizil, C. Sauper, and S. J. Taylor. Characterizing Online Public Discussions through Patterns of Participant Interactions. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW), 2018.